CCWG
Internet-Draft
Intended status: Standards Track
Expires: 31 August 2026

Y. Tian
J. Yang
W. Cheng
China Mobile
J. Wang
G. Zhang
Centec
K. Zhang
China Mobile

27 February 2026

Multi-level Congestion Response Framework with Long-haul Congestion Notification for DCI Networks
draft-tian-ccwg-long-haul-cnp-00

Abstract

This document specifies a multi-level congestion response framework and an associated Long-haul Congestion Notification Packet (Long-haul CNP) for Data Center Interconnect (DCI) wide-area network scenarios. The framework defines a graduated congestion response mechanism: lightweight ECN marking for incipient congestion and device-originated Long-haul CNP for severe or rapidly worsening congestion. Long-haul CNP packets carry explicit control instructions (e.g., rate reduction percentage, pause duration) and are sent directly by congestion-aware intermediate nodes to the traffic source via unicast, reducing feedback latency compared to receiver-mediated congestion notification. The document also specifies a multi-device collaborative suppression mechanism and BDP-adaptive dynamic threshold calculation for long-haul links. Two packet encapsulation formats are defined: an ICMPv6 extension and a RoCEv2 backward-compatible extension.

This Internet-Draft will expire on 31 August 2026.

Table of Contents

1. Introduction

RDMA over Converged Ethernet v2 (RoCEv2) is widely deployed in data center networks for high-performance computing and AI training workloads. Within a single data center, congestion control mechanisms such as DCQCN [DCQCN] and ECN-based schemes provide effective flow control. However, when RoCEv2 traffic traverses Data Center Interconnect (DCI) wide-area networks, the existing congestion notification path ("switch marks ECN, receiver generates CNP, CNP returns to sender") introduces feedback latency proportional to the WAN round-trip time, which can reach tens of milliseconds.

Recent work on Fast CNP [I-D.xiao-rtgwg-rocev2-fast-cnp] has addressed the fundamental latency issue by enabling switches to generate CNP packets directly to the sender. This document builds upon that foundation by specifying three complementary mechanisms that are particularly relevant for long-haul DCI scenarios:

1. A graduated multi-level trigger framework that distinguishes between incipient and severe congestion, avoiding unnecessary control packet generation for transient queue buildup.

2. A BDP-adaptive dynamic threshold calculation that automatically adjusts congestion detection sensitivity based on link characteristics (bandwidth, distance, RTT).

3. A multi-device collaborative suppression mechanism that coordinates congestion responses across multiple intermediate nodes on a data flow path, preventing redundant or conflicting control instructions.

Additionally, this document defines an extended packet format (Long-haul CNP) that carries explicit control instructions with quantified congestion metrics, enabling the source to perform precise rate adjustments rather than relying on generic rate reduction heuristics.

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. Terminology

RDMA: Remote Direct Memory Access.

RoCEv2: RDMA over Converged Ethernet version 2 [RoCEv2].

CNP: Congestion Notification Packet, as defined in the RoCEv2 specification. The standard CNP uses BTH Opcode 0x81.

Fast CNP: Fast Congestion Notification Packet, a switch-originated CNP as defined in [I-D.xiao-rtgwg-rocev2-fast-cnp].

Long-haul CNP: Long-haul Congestion Notification Packet, the extended packet type defined in this document, carrying explicit control instructions and congestion metrics.

ECN: Explicit Congestion Notification [RFC3168].

QP: Queue Pair, a fundamental RDMA communication abstraction.

BTH: Base Transport Header, the common header in all RoCEv2 packets.

RTT: Round-Trip Time.

PFC: Priority-based Flow Control.

DCI: Data Center Interconnect.

BDP: Bandwidth-Delay Product.

Congestion-Aware Intermediate Node: A network device deployed on the DCI path that implements the multi-level congestion monitoring and Long-haul CNP generation functions specified in this document. This may be a router, switch, or dedicated DCI gateway device.

3. Related Work and Positioning

This section describes the relationship between this document and existing congestion notification mechanisms for RoCEv2 networks.

[I-D.xiao-rtgwg-rocev2-fast-cnp] defines the Fast CNP mechanism, which enables a switch to generate a CNP packet directly to the sender when it detects congestion, without waiting for the receiver to generate the CNP. Fast CNP provides the foundational mechanism for switch-originated congestion notification.

This document extends the Fast CNP concept in three directions:

Graduated Response: While Fast CNP triggers upon detecting congestion, this document defines a two-level response where lightweight ECN marking handles incipient congestion and Long-haul CNP generation is reserved for severe or rapidly worsening conditions. This reduces control packet overhead in mildly congested scenarios.

Instructional Control: While Fast CNP advises the sender to reduce its rate, the Long-haul CNP packet format defined here carries explicit action codes (notify, pause, rate reduce, resume) with quantified parameters (e.g., reduction percentage, pause duration) and congestion metrics, enabling more precise source-side rate adjustment.

Multi-device Coordination: In DCI scenarios where a flow traverses multiple congestion-aware nodes, this document defines coordination rules to prevent duplicate or conflicting control instructions from reaching the same source.

The mechanisms defined in this document are complementary to SRv6-based congestion control approaches such as those described in [I-D.liu-spring-srv6-cc] and [I-D.hu-rtgwg-rocev2-fcn]. When used within SRv6-based DCI networks [RFC8754], the Long-haul CNP can be encapsulated within the applicable SRv6 transport framework.

RTT-based congestion control approaches such as TIMELY [TIMELY] provide an alternative signal (delay) for inferring congestion severity; the Congestion Metric field defined in this document can optionally carry RTT-derived information to complement queue-based metrics.

4. Applicability Statement

The mechanisms specified in this document are primarily designed for Data Center Interconnect (DCI) scenarios where RoCEv2 traffic traverses wide-area network paths with non-trivial propagation delay (typically RTT greater than 1 ms). In such environments, the receiver-mediated CNP feedback path introduces significant latency, and the BDP-adaptive threshold mechanism provides meaningful dynamic range.

For intra-data-center deployments where RTT is sub-millisecond and paths traverse few hops, the standard ECN/CNP or Fast CNP mechanisms are typically sufficient, and the additional complexity of the multi-level framework may not be warranted.

The multi-device collaborative suppression mechanism is most beneficial when data flows traverse two or more congestion-aware intermediate nodes, which is common in multi-hop DCI topologies.

Regarding IP version applicability: the ICMPv6 packet format defined in Section 6.1 is applicable only to IPv6 network deployments. For DCI environments that operate over IPv4, implementations MUST use the RoCEv2 backward-compatible extension format defined in Section 6.2. A future document may define an ICMPv4-based format if there is sufficient demand for ICMP-based Long-haul CNP in IPv4-only deployments.

5. Protocol Specification

5.1. Architecture Overview

The following diagram illustrates a typical DCI topology where this mechanism operates:

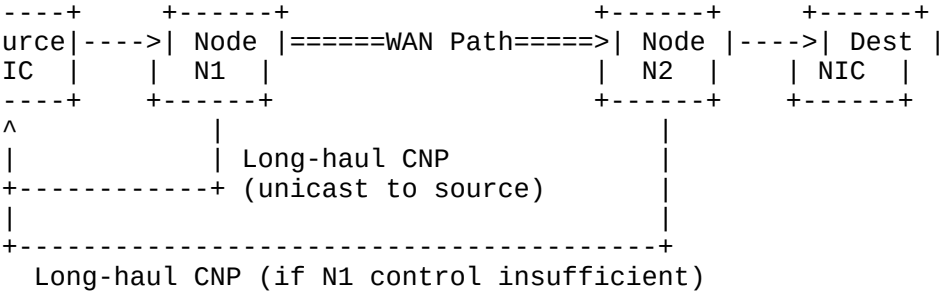

Figure 1: DCI Network Topology with Congestion-Aware Nodes

The mechanism operates as follows:

1. Congestion-aware intermediate nodes (N1, N2) learn flow state by inspecting traversing RoCEv2 data packets.

2. Each node continuously monitors egress queue status using multiple congestion indicators.

3. When incipient congestion is detected (queue depth exceeds K_min), the node applies ECN marking to traversing data packets (first-level response).

4. When severe congestion is detected (any second-level condition is met), the node generates a Long-haul CNP packet and sends it via unicast to the traffic source (second-level response).

5. The source parses the Long-haul CNP and adjusts the sending rate of the indicated QP according to the control instruction.

6. If multiple nodes detect congestion for the same flow, the collaborative suppression mechanism coordinates their responses.

5.2. Flow Table Learning and Maintenance

A congestion-aware intermediate node MUST parse the Base Transport Header (BTH) of traversing RoCEv2 data packets and extract the following flow identification information: Source IP Address, Destination IP Address, Source QP Number, and Destination QP Number.

The node SHOULD maintain a flow table with one entry per unique four-tuple (Source IP, Destination IP, Source QP, Destination QP). Flow table entries MUST be updated upon each matching packet observation. Entries MAY be subject to an aging timer; when no matching packets are observed within the configured aging period, the entry SHOULD be removed.

The flow table is used to construct Long-haul CNP packets with the correct addressing information when congestion is detected.

5.3. Multi-level Congestion Monitoring

5.3.1. Monitoring Metrics

Congestion-aware intermediate nodes MUST continuously monitor the following metrics on each egress port:

Queue Depth (QD): The instantaneous volume of data buffered in the egress queue, measured in bytes or cells.

ECN Marking Rate (EMR): The fraction of data packets marked with ECN per unit time on the egress port.

Queue Growth Rate (QGR): The rate of change of queue depth over a configurable measurement interval, indicating whether congestion is building or subsiding.

The node SHOULD also maintain an estimate of the link round-trip time (RTT_est) for each egress port, which MAY be obtained through control plane configuration, active probing, or receiver reporting.

5.3.2. BDP-Adaptive Dynamic Threshold Calculation

To account for the wide variation in link characteristics across DCI paths, the upper queue depth threshold K_max MUST be dynamically calculated based on the Bandwidth-Delay Product (BDP) of the link:

```
K_max = max(K_base, alpha * R_port * RTT_est / 8)
```

R_port: The egress port rate in bits per second.

RTT_est: The estimated round-trip time in seconds.

K_base: A baseline queue threshold providing a minimum sensitivity floor for short-distance or low-speed links. The RECOMMENDED default value is implementation- specific but SHOULD correspond to at least one maximum- sized RoCEv2 frame.

alpha: An adjustment coefficient. The RECOMMENDED default value is 1.0. Implementations MAY allow this to be configured per-port or per-link.

The minimum threshold K_min, used for first-level ECN marking, MUST be configured to a value less than K_max. A RECOMMENDED default is K_min = K_max / 2.

Implementations SHOULD recalculate K_max periodically or upon RTT_est changes, to adapt to evolving link conditions.

5.4. Multi-level Trigger Response

This document defines two levels of congestion response:

First-level response (ECN marking): When the queue depth QD exceeds K_min, the node MUST apply ECN marking (setting the CE codepoint per [RFC3168]) to traversing data packets. This activates the standard end-to-end ECN/CNP feedback loop and does not generate any additional control packets. First-level response operates as a lightweight, low-overhead mechanism for handling transient or mild congestion.

Second-level response (Long-haul CNP generation): The node MUST trigger Long-haul CNP generation when any of the following conditions is satisfied: (a) Queue depth QD exceeds K_max; (b) ECN marking rate EMR exceeds a configured threshold V_ecn; (c) Queue growth rate QGR exceeds a configured threshold V_growth. Second-level response is intended for severe or rapidly worsening congestion where the first-level ECN feedback loop cannot respond quickly enough due to WAN path latency.

5.5. Long-haul CNP Generation and Transmission

When a second-level trigger condition is met, the congestion-aware intermediate node MUST perform the following procedure:

1. Identify the data flow(s) contributing to the congested queue. The node SHOULD select the flow(s) with the highest contribution to the queue occupancy when multiple flows share the queue.

2. For each selected flow, look up the corresponding flow table entry to obtain the source IP address and source QP number.

3. Determine the appropriate action instruction (Action Flags) and parameter values based on the current congestion severity (see Section 5.5.1).

4. Construct a Long-haul CNP packet as specified in Section 6, populating all required fields.

5. Send the packet via unicast to the source IP address of the identified flow.

The node MUST rate-limit Long-haul CNP generation to prevent excessive control traffic. The RECOMMENDED minimum interval between consecutive Long-haul CNP packets for the same flow is one estimated RTT (RTT_est).

5.5.1. Action Instruction Determination

The Action Flags field in the Long-haul CNP packet encodes the specific control action requested of the source. The following guidelines apply:

Notify (00): Informs the source that congestion exists. The source SHOULD apply its default congestion response algorithm. This action is used when congestion is detected but not yet severe.

Pause (01): Instructs the source to temporarily halt transmission on the indicated QP for the duration specified in the Parameter field (in microseconds). This action SHOULD be used only for severe congestion where rate reduction alone is insufficient.

Rate Reduce (10): Instructs the source to reduce its sending rate on the indicated QP by the percentage specified in the Parameter field. For example, a Parameter value of 30 indicates a 30% rate reduction.

Resume (11): Instructs the source that congestion has subsided and

the source MAY restore its sending rate on the indicated QP. The Parameter field specifies the permitted rate recovery percentage relative to the rate prior to the last congestion action. For example, a Parameter value of 50 indicates the source MAY increase its rate by 50% of the reduction that was previously applied. A Parameter value of 0 indicates unconditional resume to the original rate. This action is generated when a congestion-aware intermediate node observes that queue depth has fallen below K_min for a sustained period (RECOMMENDED: at least 1 * RTT_est).

5.6. Source Behavior upon Receiving Long-haul CNP

Upon receiving a Long-haul CNP packet, the source MUST:

1. Validate the packet by checking that the Source QP Number matches a locally active QP and that the source IP of the Long-haul CNP belongs to a known intermediate node (see Section 9).

2. Parse the Action Flags and Parameter fields.

3. Apply the indicated action to the corresponding QP’s sending rate or transmission state.

The source SHOULD maintain a per-QP timer. If no new Long-haul CNP packet is received for the same QP within a configurable recovery interval (RECOMMENDED: 2 * RTT_est), the source SHOULD gradually increase its sending rate using an additive-increase algorithm until normal rate is restored or a new Long-haul CNP is received.

Upon receiving a Resume action, the source SHOULD increase its sending rate by the percentage indicated in the Parameter field. The source MAY combine the timer-based recovery mechanism with explicit Resume actions: when a Resume is received, the source applies the indicated rate increase immediately rather than waiting for the recovery timer.

If a source receives a Long-haul CNP but does not support the Long-haul CNP format, it MUST silently discard the packet (ICMPv6 format) or process it as a standard CNP (RoCEv2 format). This ensures backward compatibility with sources that only support standard CNP. See Section 8 for details.

5.7. Multi-device Collaborative Congestion Suppression

When a data flow traverses multiple congestion-aware intermediate nodes, uncoordinated Long-haul CNP generation can result in duplicate or conflicting control instructions reaching the source. This section specifies coordination rules to mitigate this issue.

Upstream Priority: When congestion occurs, the node closest to the congestion point (in the upstream direction toward the source) SHOULD respond first. Because this node's Long-haul CNP has the shortest path to the source, it achieves the fastest feedback.

Downstream Deferral: A downstream congestion-aware node that detects congestion for a given flow SHOULD check whether a Long-haul CNP for the same flow has recently been generated by an upstream node. This can be inferred by observing a reduction in the flow's arrival rate within a configurable observation window (RECOMMENDED: 1 * RTT_est). If such a reduction is observed:

a. The downstream node SHOULD temporarily suppress Long-haul CNP generation for that flow, to avoid sending duplicate instructions.

b. If the downstream node's congestion metrics continue to worsen despite the observation window (i.e., the upstream control has not been effective), the downstream node MUST generate a Long-haul CNP with a higher Congestion Level value and a stricter action instruction (e.g., escalating from Rate Reduce to Pause).

5.8. Dynamic Parameter Adjustment

Congestion-aware intermediate nodes MAY dynamically adjust the threshold parameters (K_min, K_max, V_ecn, V_growth) and rate-limiting intervals based on observed traffic characteristics such as average queue occupancy, traffic burstiness patterns, and link utilization history. The specific algorithms for such adjustment are implementation-dependent and outside the scope of this document.

6. Packet Formats

This document defines two Long-haul CNP packet formats. Implementations MUST support at least one format and SHOULD indicate the supported format(s) through out-of-band configuration or capability exchange.

6.1. ICMPv6 Extension Format

This format encapsulates the Long-haul CNP as a new ICMPv6 informational message type [RFC4443]. Because the Long-haul CNP is a new ICMPv6 message type with a fully defined fixed-length body (no variable-length "original datagram" field), the length ambiguity problem addressed by [RFC4884] does not apply. However, the Optional Extension Objects defined in this section adopt the extension structure format from [RFC4884] (Extension Header with Version and

Checksum, and Extension Objects with Class-Num, C-Type, Length) for consistency with IETF ICMP extension conventions and to enable reuse of existing ICMP extension parsing implementations.

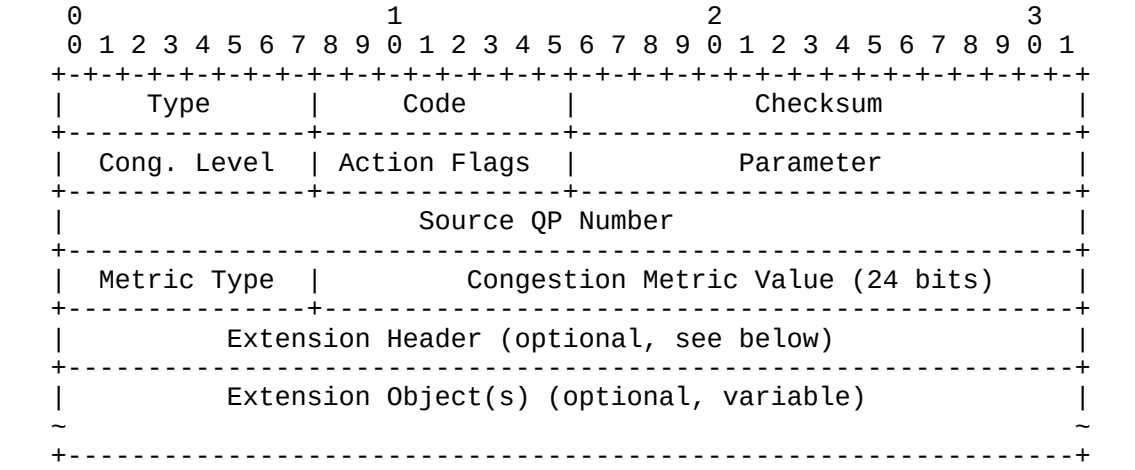

Figure 2: Long-haul CNP in ICMPv6 Format

The fixed-length body of the Long-haul CNP ICMPv6 message is 12 octets (3 x 32-bit words), comprising the fields from Congestion Level through Congestion Metric. This fixed length is known to all implementations, so the boundary between the fixed body and any extension structure is unambiguous.

Type (8 bits): ICMPv6 message type. A new value is to be assigned by IANA from the informational message range (128-255). See Section 10.

Code (8 bits): Message sub-type. Value 0 indicates a flow-level Long-haul CNP. Other values are reserved for future definition. See Section 10.2 for the registration policy.

Checksum (16 bits): Standard ICMPv6 checksum as specified in [RFC4443].

Congestion Level (8 bits): An unsigned integer indicating the severity of congestion, where 0 indicates no congestion and 255 indicates maximum severity. This value is determined by the generating node based on its local congestion assessment.

Action Flags (8 bits): The upper 2 bits encode the primary action:

00 = Notify, 01 = Pause, 10 = Rate Reduce, 11 = Resume. The lower 6 bits are reserved and MUST be set to zero by senders. Receivers MUST ignore the reserved bits.

Parameter (16 bits): Semantics depend on Action Flags. For Rate Reduce: rate reduction percentage (0-100). For Pause: pause duration in microseconds. For Resume: rate recovery percentage (0-100), where 0 indicates unconditional resume to the original rate. For Notify: unused, MUST be set to zero.

Source QP Number (32 bits): The QP number at the traffic source that should apply the indicated action.

Metric Type (8 bits): Identifies the type of the Congestion Metric Value field. Defined values are: 0 = Unspecified (implementation-defined semantics, for backward compatibility), 1 = Queue Depth in kilobytes, 2 = Queue Growth Rate in kilobytes per millisecond, 3 = ECN Marking Rate as a percentage (0-100), 4 = RTT-based metric in microseconds, 5-253 = Unassigned (see Section 10.4), 254-255 = Experimental. When Metric Type is 0, receivers SHOULD treat the Congestion Metric Value as opaque context.

Congestion Metric Value (24 bits): An unsigned integer whose semantics are determined by the Metric Type field. This field provides additional context for source-side decision-making. When the generating node does not wish to disclose queue state information, both Metric Type and Congestion Metric Value MUST be set to zero.

6.1.1. Optional Extension Objects

Zero or more extension objects MAY follow the fixed-length body. When extension objects are present, they MUST be preceded by an Extension Header as defined in Section 3 of [RFC4884], formatted as follows:

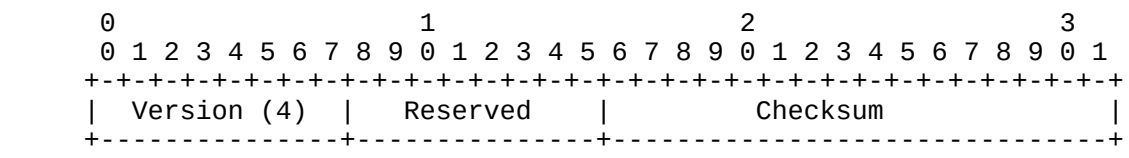

Figure 3: Extension Header Format

Version (4 bits): MUST be set to 2, as specified in Section 3 of [RFC4884].

Reserved (12 bits): MUST be set to zero.

Checksum (16 bits): The one's complement of the one's complement sum of the Extension Header and all Extension Objects, computed as specified in Section 3 of [RFC4884].

Each Extension Object following the Extension Header uses the object header format defined in Section 4 of [RFC4884]:

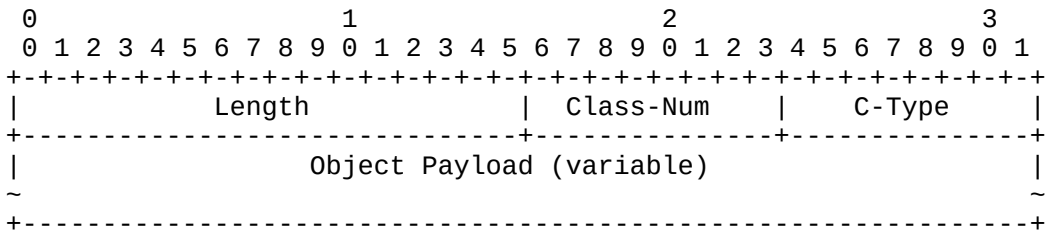

Figure 4: Extension Object Format

Length (16 bits): Total length of the Extension Object in octets, including the 4-octet object header. As specified in [RFC4884], each Extension Object MUST be zero-padded to a 4-octet boundary. The Length field indicates the actual (unpadded) length; receivers MUST use the padded length when advancing to the next Extension Object.

Class-Num (8 bits): Identifies the class of the Extension Object. A new Class-Num is to be assigned by IANA (see Section 10).

C-Type (8 bits): Identifies the sub-type within the class. Defined C-Types include: 0 = Reserved, 1 = Timestamp (8-octet NTP timestamp), 2 = Device Identifier (variable-length UTF-8 string, padded to 4-octet boundary), 3 = Path Identifier (variable-length opaque value, padded to 4-octet boundary).

When no extension objects are present, the Extension Header MUST be omitted entirely. Receivers determine the presence of extension objects by checking whether the ICMPv6 message length exceeds the fixed body length (12 octets beyond the standard 4-octet ICMPv6 header).

6.2. RoCEv2 Backward-Compatible Extension Format

This format achieves backward compatibility by reusing the standard CNP BTH Opcode (0x81) and extending the packet through a reserved bit in the BTH. This approach avoids the need for IETF to request a new Opcode from the InfiniBand Trade Association (IBTA), while ensuring that legacy endpoints that do not support Long-haul CNP will still process the packet as a standard CNP and apply their default rate

reduction behavior.

6.2.1. Extension Present (E) Bit Definition

In the standard RoCEv2 BTH, the 6-bit field immediately following the BECN (Backward Explicit Congestion Notification) bit is reserved and MUST be set to zero per the current RoCEv2 specification. This document proposes to the IBTA the definition of the most significant bit of this 6-bit reserved field as the Extension Present (E) bit:

E = 0: Standard CNP. No extension fields follow the BTH. The packet is processed as a conventional RoCEv2 CNP.

E = 1: Long-haul CNP. Extension fields carrying congestion control instructions follow the BTH. The packet retains all standard CNP BTH field values (Opcode=0x81, BECN=1) and is fully parseable as a standard CNP by legacy endpoints.

The E bit definition is a proposal for IBTA consideration. This bit resides within a reserved field that is under IBTA governance, and formal allocation requires IBTA approval. Prior to such approval, implementations MUST NOT deploy this format in environments where non-participating endpoints or intermediate nodes may be present, as legacy devices that validate the reserved field as zero may reject packets with E=1. See Section 10.5 for further coordination details.

6.2.2. Packet Layout

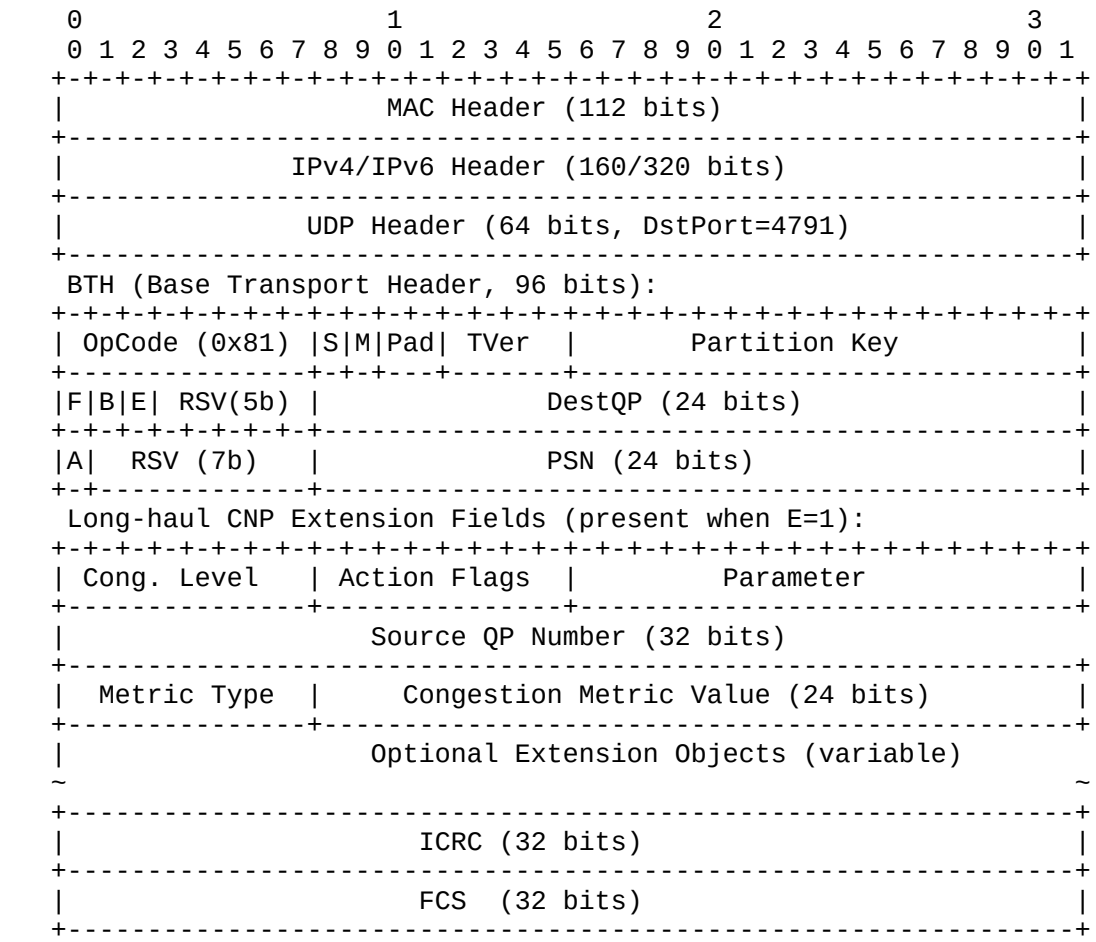

Figure 5: Long-haul CNP in RoCEv2 Extension Format

6.2.3. BTH Field Values for Long-haul CNP

When generating a Long-haul CNP in RoCEv2 format, the congestion-aware intermediate node MUST set the BTH fields as follows:

OpCode (8 bits): MUST be set to 0x81 (binary 10000001), the standard CNP opcode as defined in the RoCEv2 specification [RoCEv2].

Solicited Event - SE (1 bit): MUST be set to 0.

MigReq - M (1 bit): MUST be set to 0.

Pad Count - PadCnt (2 bits): MUST be set to 0.

Transport Header Version - TVer (4 bits): MUST be set to 0x0.

Partition Key - P_KEY (16 bits): MUST be set to 0xFFFF, or the partition key value of the target flow if partition-aware operation is required.

FECN - F (1 bit): MUST be set to 0.

BECN - B (1 bit): MUST be set to 1, indicating backward congestion notification. This is the standard CNP BECN setting.

Extension Present - E (1 bit): MUST be set to 1 for Long-haul CNP. This bit indicates that Long-haul CNP extension fields follow the BTH.

Reserved (5 bits): MUST be set to 0.

DestQP (24 bits): The destination QP number at the source to be controlled. In standard CNP semantics, this identifies the QP that should reduce its sending rate.

Acknowledge Request - A (1 bit): MUST be set to 0.

Reserved (7 bits): MUST be set to 0.

PSN (24 bits): MUST be set to 0, consistent with standard CNP behavior.

6.2.4. Extension Fields

The extension fields immediately follow the BTH when E=1. Their encoding is consistent with the ICMPv6 format:

Congestion Level (8 bits): Encoding consistent with Section 6.1.

Action Flags (8 bits): Encoding consistent with Section 6.1.

Parameter (16 bits): Encoding consistent with Section 6.1.

Source QP Number (32 bits): Source-side QP number for precise flow identification. This field complements the DestQP in the BTH to form the complete QP pair identification.

Metric Type (8 bits): Encoding consistent with Section 6.1.

Congestion Metric Value (24 bits): Encoding consistent with Section 6.1.

Optional Extension Objects (variable): Zero or more Extension Objects, using the same Extension Header and Extension Object format as defined in Section 6.1.1. When present, the Extension Header MUST precede the first Extension Object.

ICRC (32 bits): Invariant CRC as specified in the RoCEv2 specification. The ICRC computation MUST include the extension fields and any Extension Objects. See Section 6.2.5 for ICRC compatibility analysis.

FCS (32 bits): Ethernet Frame Check Sequence.

6.2.5. ICRC Compatibility Analysis

In the RoCEv2 specification, the ICRC is computed over all bytes from the beginning of the BTH to the byte immediately preceding the ICRC field itself (with certain IP and UDP header fields replaced by defined values). When a Long-haul CNP is constructed with E=1, the extension fields and any Optional Extension Objects are placed between the BTH and the ICRC field. Therefore, the ICRC computation naturally covers the extension data.

A legacy RNIC that receives a Long-haul CNP will compute the ICRC over the same byte range (BTH through the byte preceding the ICRC field). Because the ICRC is always located at a fixed offset from the end of the Ethernet frame (immediately before the FCS), the legacy RNIC will include the extension fields in its ICRC computation even though it does not parse them. Consequently, the ICRC verification will succeed on legacy endpoints, and no ICRC mismatch will occur due to the presence of extension fields.

7. Operational Example

Consider a DCI path: Source (DC-A) -> N1 -> N2 -> Dest (DC-B), where N1 and N2 are congestion-aware intermediate nodes, and the WAN RTT is 10 ms.

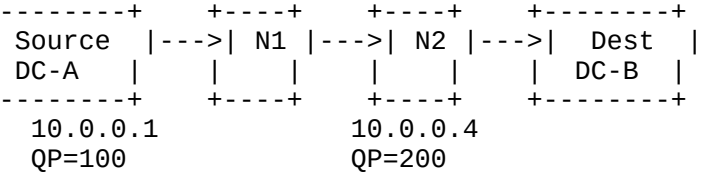

Figure 6: Operational Example Topology

1. N1 learns the flow: {Src=10.0.0.1, Dst=10.0.0.4, SrcQP=100, DstQP=200}.

2. N1 calculates K_max for its egress port (100 Gbps link): K_max = max(64KB, 1.0 * 100e9 * 0.010 / 8) = max(64KB, 125MB) = 125 MB. K_min = 62.5 MB.

3. Queue depth at N1 reaches 70 MB (exceeds K_min). N1 applies ECN marking to traversing packets (first-level response).

4. Queue depth continues to grow to 130 MB (exceeds K_max). N1 generates a Long-haul CNP: {Action=Rate Reduce, Parameter=30, SourceQP=100, CongLevel=180, MetricType=1 (Queue Depth in KB), MetricValue=130000} and unicasts it to 10.0.0.1. In RoCEv2 format, the BTH uses Opcode=0x81, BECN=1, E=1, DestQP=100.

5. Source receives the Long-haul CNP within approximately 5 ms (half RTT). If the source supports Long-haul CNP, it checks E=1 and parses the extension fields, reducing QP 100’s sending rate by 30%. If the source is a legacy RNIC that does not recognize the E bit, it processes the packet as a standard CNP and applies its default rate reduction algorithm (see Section 8 for frame length considerations).

6. N2 also detects mild congestion but observes that the flow arrival rate from N1 has decreased. N2 defers Long-haul CNP generation.

7. After 20 ms without new Long-haul CNP, the source begins additive rate recovery on QP 100.

8. N1 observes queue depth has fallen below K_min for more than 10 ms (1 * RTT_est). N1 generates a Long-haul CNP: {Action=Resume, Parameter=50, SourceQP=100, CongLevel=20, MetricType=1, MetricValue=30000}. The source increases QP 100’s rate by 50% of the previously applied reduction.

8. Backward Compatibility

The Long-haul CNP mechanism is designed for incremental deployment:

Non-supporting intermediate nodes: Intermediate nodes that do not implement this specification simply forward data packets without generating Long-haul CNP. The standard ECN/CNP feedback loop continues to operate as the baseline congestion control mechanism.

Non-supporting sources (ICMPv6 format): If a source receives an ICMPv6 Long-haul CNP but does not recognize the Type value, it MUST process it according to standard ICMPv6 unknown-type handling rules [RFC4443]. For informational messages (Type values 128-255), the unknown message is silently discarded.

Non-supporting sources (RoCEv2 format): Because the Long-haul CNP in RoCEv2 format uses the standard CNP Opcode (0x81) with all mandatory BTH fields set to their standard CNP values, a legacy RNIC that does not recognize the E bit will process the packet as a standard CNP. The legacy RNIC will ignore the extension fields (which appear after the expected CNP boundary) and apply its default rate reduction behavior. This provides a graceful degradation path: precise control for supporting endpoints, and standard CNP rate reduction for legacy endpoints.

However, the Long-haul CNP frame is larger than a standard CNP frame due to the extension fields (at least 12 additional octets for the base extension, plus any Optional Extension Objects). Legacy RNIC implementations that perform strict frame length validation against the expected standard CNP size may reject the Long-haul CNP packet. Deployments SHOULD verify that legacy endpoints in the network tolerate CNP frames with additional trailing data beyond the standard BTH before enabling the RoCEv2 Long-haul CNP format. In environments where legacy endpoints are known to perform strict length validation, the ICMPv6 format SHOULD be used instead, or all endpoints should be upgraded to support the Long-haul CNP extension.

Mixed deployment: In networks where only some intermediate nodes support this specification, the supporting nodes generate Long-haul CNP while non-supporting nodes rely on ECN marking alone. The multi-level framework degrades gracefully to standard ECN/CNP behavior in portions of the path without Long-haul CNP capability.

9. Security Considerations

Long-haul CNP packets carry control instructions that directly affect the source’s sending behavior. The following security considerations apply:

Packet Forgery: A malicious entity could forge Long-haul CNP packets to cause a source to reduce its rate or pause transmission, resulting in denial of service. To mitigate this, sources SHOULD validate that the IP source address of received Long-haul CNP packets belongs to a configured set of known congestion-aware intermediate node addresses. For the ICMPv6 format, IPsec Authentication Header (AH) MAY be used to provide packet authentication. For the RoCEv2 format, the ICRC provides integrity protection but not authentication; additional authentication mechanisms are RECOMMENDED in security-sensitive deployments.

Amplification: A single congested packet could potentially trigger

Long-haul CNP generation targeting multiple sources. Congestion-aware intermediate nodes MUST rate-limit Long-haul CNP generation on a per-flow basis (RECOMMENDED minimum interval: RTT_est per flow) and MUST impose a global rate limit on total Long-haul CNP output per port.

Information Disclosure: The Congestion Metric Value field reveals internal queue state information. In deployments where this is considered sensitive, both the Metric Type and Congestion Metric Value fields MUST be set to zero while still providing actionable control via the Action Flags and Parameter fields.

Reserved Bit Manipulation (RoCEv2 format): A malicious entity that can modify packets in transit could set the E bit on standard CNP packets and append forged extension fields. The ICRC field provides integrity protection against in-transit modification for RoCEv2 packets. Additionally, sources SHOULD validate that the combination of extension field values is consistent before applying control actions.

Resume Action Abuse: A malicious entity could forge Long-haul CNP packets with the Resume action to cause a source to prematurely increase its sending rate during actual congestion. The source address validation described under Packet Forgery above mitigates this risk. Additionally, sources SHOULD apply a maximum rate increase cap per Resume action to limit the impact of any single forged Resume instruction.

10. IANA Considerations

10.1. ICMPv6 Type Allocation

This document requests IANA to allocate a new value from the "ICMPv6 'type' Numbers" registry for the Long-haul CNP message type. The value SHOULD be allocated from the informational message range (128-255).

Type	Name	Reference
TBD1	Long-haul Congestion Notification	[This Document]

Table 1

10.2. ICMPv6 Code Values

This document requests IANA to create a sub-registry titled "Long-haul Congestion Notification Code Values" under the "ICMPv6 'type' Numbers" registry, for Code values associated with the ICMPv6 Type allocated in Section 10.1. The initial contents of this sub-registry are:

Code	Name	Reference
0	Flow-level Long-haul CNP	[This Document]
1-253	Unassigned	
254-255	Experimental	[This Document]

Table 2

New Code values in the range 1-253 are to be assigned via Specification Required policy [RFC8126].

10.3. ICMP Extension Object Class-Num

This document requests IANA to allocate a new Class-Num value from the "ICMP Extension Object Classes and Class Sub-types" registry established by [RFC4884].

Class-Num	Class Name	Reference
TBD2	Long-haul CNP Extension	[This Document]

Table 3

Within this Class-Num, the following C-Type values are defined:

C-Type	Name	Reference
0	Reserved	[This Document]
1	Timestamp	[This Document]
2	Device Identifier	[This Document]
3	Path Identifier	[This Document]
4-253	Unassigned	
254-255	Experimental	[This Document]

Table 4

New C-Type values in the range 4-253 are to be assigned via Specification Required policy [RFC8126].

10.4. Congestion Metric Type Values

This document requests IANA to create a new registry titled "Long-haul CNP Congestion Metric Type Values". The initial contents of this registry are:

Value	Name	Unit	Reference
0	Unspecified	N/A	[This Document]
1	Queue Depth	Kilobytes	[This Document]
2	Queue Growth Rate	Kilobytes/ms	[This Document]
3	ECN Marking Rate	Percentage (0-100)	[This Document]
4	RTT-based Metric	Microseconds	[This Document]
5-253	Unassigned		
254-255	Experimental		[This Document]

Table 5

New values in the range 5-253 are to be assigned via Specification Required policy [RFC8126].

10.5. RoCEv2 Reserved Bit Coordination

The RoCEv2 Long-haul CNP format defined in this document proposes the use of the most significant bit of the 6-bit reserved field following the BECN bit in the BTH as an Extension Present (E) bit. The RoCEv2 BTH format is defined by the InfiniBand Trade Association (IBTA), and the reserved field is under IBTA governance.

This document respectfully requests that the IBTA consider allocating this bit as the "Long-haul Extension Present" indicator for CNP packets (Opcode 0x81). The E bit definition specified in this document is a proposal intended to facilitate IBTA review; it does not constitute a unilateral allocation by the IETF of IBTA-governed protocol space.

Until IBTA formally approves this allocation, implementations of the RoCEv2 format defined in this document are considered experimental and MUST only be deployed in controlled environments where all endpoints and intermediate nodes are known to support this extension. Specifically, implementations MUST NOT send Long-haul CNP packets in RoCEv2 format to endpoints that have not been explicitly configured or negotiated to accept them.

11. References

11.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, DOI 10.17487/RFC3168, September 2001, <https://www.rfc-editor.org/info/rfc3168>.

[RFC4443] Conta, A., Deering, S., and M. Gupta, Ed., "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", STD 89, RFC 4443, DOI 10.17487/RFC4443, March 2006, <https://www.rfc-editor.org/info/rfc4443>.

[RFC4884] Bonica, R., Gan, D., Tappan, D., and C. Pignataro, "Extended ICMP to Support Multi-Part Messages", RFC 4884, DOI 10.17487/RFC4884, April 2007, <https://www.rfc-editor.org/info/rfc4884>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, June 2017, <https://www.rfc-editor.org/rfc/rfc8126>.

11.2. Informative References

[I-D.xiao-rtgwg-rocev2-fast-cnp] Min, X., Li, H., Zhang, K., Cheng, W., Yang, J., and K. Zhang, "Fast Congestion Notification Packet (CNP) in RoCEv2 Networks", Work in Progress, Internet-Draft, draft-xiao-rtgwg-rocev2-fast-cnp-04, December 2025, <https://datatracker.ietf.org/doc/html/draft-xiao-rtgwg-rocev2-fast-cnp-04>.

[I-D.liu-spring-srv6-cc] Liu, Y. and H. Shi, "Congestion Control Based on SRv6 Path", Work in Progress, Internet-Draft, draft-liu-spring-srv6-cc-01, July 2025, <https://datatracker.ietf.org/doc/html/draft-liu-spring-srv6-cc-01>.

[I-D.hu-rtgwg-rocev2-fcn] Hu, Z., Zhu, Y., and X. Geng, "Fast congestion notification for distributed RoCEv2 network based on SRv6", Work in Progress, Internet-Draft, draft-hu-rtgwg-rocev2-fcn-00, March 2025, <https://datatracker.ietf.org/doc/html/draft-hu-rtgwg-rocev2-fcn-00>.

[DCQCN] Zhu, Y., "Congestion Control for Large-Scale RDMA Deployments", ACM SIGCOMM, 2015.

[TIMELY] Mittal, R., "TIMELY: RTT-based Congestion Control for the Datacenter", ACM SIGCOMM, 2015.

[RoCEv2] InfiniBand Trade Association, "Supplement to InfiniBand Architecture Specification Volume 1 Release 1.2.1 - Annex A17: RoCEv2", 2014.

[RFC8754] Filsfils, C., Ed., Dukes, D., Ed., Previdi, S., Leddy, J., Matsushima, S., and D. Voyer, "IPv6 Segment Routing Header (SRH)", RFC 8754, DOI 10.17487/RFC8754, March 2020, <https://www.rfc-editor.org/info/rfc8754>.

Authors' Addresses

Yuchi Tian
China Mobile
Beijing
100053
China
Email: tianyuchi@chinamobile.com

Jin Yang
China Mobile
Beijing
100053
China
Email: yangjin@chinamobile.com

Weiqiang Cheng
China Mobile
Beijing
100053
China
Email: chengweiqiang@chinamobile.com

Junjie Wang
Centec
Suzhou
215000
China
Email: wangjj@centec.com

Guoying Zhang
Centec
Suzhou
215000
China
Email: zhanggy@centec.com

Kan Zhang
China Mobile
Beijing
100053
China
Email: zhangkan@chinamobile.com